Planning the Industrial Estate Area for Comparison Two Methods

Edy Fradinata, ST, MT
Industrial Engineering Department of
Escambia Makah University
Banda Aceh, Indonesia
edinata69@gmail.com

Abstract— Analysis cluster method is a popular science in the knowledge of partitioning data in the areas of engineering, medicine, economics and statistics are used to perform machine learning, data mining and other data grouping. This method is often used because it is very easy and able to partition data quickly on large data sets. Cluster analysis plays an important role in classifying the object, depending on the application; object can be in the form of signals, customers, patients, news, and other plants. This technique is a nonparametric technique which is very much applicable in real cases. Clustering techniques can be grouped into two major classes: partitioning the cluster (K-means) and hierarchical cluster. There are two kinds of clustering techniques are often used. The first is the K-means and the second is a hierarchical clustering. In this research uses of investment data as the primary data source will be analyzed by comparing the two algorithm methods, both of the above algorithm to find the final solution using high-level programming language. In this research result that quite the same in application data to planning the industrial estate area.

Keywords—Clustering; K-means; Partitioning cluster; hierarchical cluster, non parametric.

I. INTRODUCTION

Clustering techniques have been well known and widely used studies in engineering and social knowledge. The main purpose of Clustering is a grouping of a large number of methods from data / object into a cluster (group) thus in each Clustering will contain data which as closely as possible with identify using distance (Euclidian) [13],[10]. Partitioning cluster is one of the most commonly used is the k-means where simplicity and speed in classifying large dataset and then hierarchical clustering of the clustering hierarchy. For k-means, k indicates the number of clusters in which the value of k specified by the user or the user. For cases where there is no consideration of a competent expert or expert in their field, the value of k will be easily determined. But it often happens that the value of k is to be determined by looking at the data. Hierarchical clustering using different partitioning approaches such as K-means clustering. In clustering hierarchy initially each data point is a cluster and then each data point is calculated similarity with the other data points in which the two most similar data points will be merged into one cluster. This is done repeatedly to form the next cluster until the cluster will eventually get one, depending on the user to a point where clustering will stop in accordance with the desired number of clusters, to see the similarities and the lack of resemblance to the use of distance, if the distance between the two points greater then both the point is not similar in this study will use the distance as the formation of clusters and industrial data as the primary data to be processed and then compared [13].

II. CLUSTERING

A. Hierarchical Clustering

In hierarchical cluster compute the distance of each object with any other object. Furthermore, we will find a partner who has a close object distances. So that each object will be paired with one object to another object or group that has the closest distance. The Steps - that must be done in this clustering is as follows:

- Categorize each object to the group / its cluster.
- Find the most similar pair to put in the same cluster by looking at the data in a matrix of similarity (resemblance).
- Combine two objects into a single cluster
- Repeat until the remaining one cluster

1. Similarities and dissimilarities

There is a kind practical way that could be doing in combining data from two or more objects into one cluster typically used similarities size or dissimilarities. The more similar higher chance of objects grouped in a single cluster. Conversely the lower odds do not like to be grouped in a cluster, to measure the similarity and dissimilarity between the data / object can be used multiple sizes. To be used cosine similarity measure, covariance and correlation. As for the size

of the distance, dissimilarities can be used as a means of identification. In a measure of similarity, the greater value means more similar. In contrast to the lack of resemblance of its show they tend the value increasingly similar .

• Cosine between two points x and y is defined as

$$\cos\theta = \frac{x^T y}{\|x\|\|y\|} \qquad (1)$$

Where $\|x\|$ defined as :

$$\sqrt{\sum_{i=1}^{n} x_i^2}$$

Where x is the first data and second data y

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y}) \qquad 2)$$

• The maximum distance between elements in the clusters the distance between two clusters is defined as:

$$d(A,B) = \max_{x \in A, y \in B}\{S_{xy}\} \qquad (3)$$

Where Sx is the distance the two data x and y each from cluster A and B

• The minimum distance between elements of each cluster (single linkage cluster)

$$d(A,B) = \min_{x \in A, y \in B}\{S_{xy}\} \qquad (4)$$

$$d(A,B) = \max_{x \in A, y \in B}\{S_{xy}\} \qquad (5)$$

• The minimum distance between elements of each cluaster
(single linkage cluster

$$d(A,B) = \min_{x \in A, y \in B}\{S_{xy}\} \qquad (6)$$

Where is the distance S_xy two data x and y of each of the clusters A and B.

$$d(A,B) = \min_{x \in A, y \in B}\{S_{xy}\} \qquad (7)$$

• The average distance between cluster.

$$d(A,B) = \frac{1}{n_A n_B}\sum_{x\in A}\sum_{y\in B} s\{x,y\} \qquad (8)$$

where n_A and n_B are the amount of the data in set A dan B

• Centroid linkage with this way the distance between two cluster can be drawn as follow :

$$d(A,B) = s(\bar{x},\bar{y}) \qquad (9)$$

Where : $\bar{x} = \frac{1}{n_A}\sum_{x\in A} x$ dan $\bar{x} = \frac{1}{n_A}\sum_{x\in A} y$

• Ward linkage

$$d(A,B) = \frac{n_A n_B s^2{}_{AB}}{n_A + n_B} \ldots\ldots\ldots\ldots(10)$$

where s_{AB}^2 the distance between cluster A and B using centroid linkage

2. Dendogram

Cluster tree or dendogram showing the sequence of how objects are grouped in clusters. X-axis shows the number of objects and the y-axis shows the distance between the object / cluster. Determination of the cluster tree can be done by cutting the cluster tree at a certain height it will be closely related to the grouping of objects that will determine the distribution of the cluster and its members. It can be described as follows [13]:

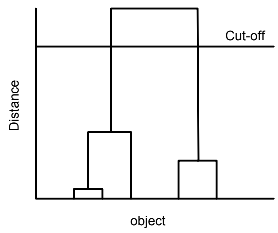

Fig.1. Cluster tree, from the bottom showing how 5th object be grouping

B. K-Means clustering

From some of the most simple clustering techniques and is commonly known *k*-means clustering. In this technique want to classify objects into k groups or clusters, to perform clustering, the value of *k* must be determined beforehand. In common the user or users already have the initial information about the object being studied; including how the most appropriate number of clusters. In detail, can be used to classify an object the size of our dissimilarities. Dissimilarities can be translated into the concept of distance. If the distance of two objects or data points are close enough, then the two objects are similar, the closer resemblance. *K*-means clustering algorithm has the following steps :

1. Select the number of clusters *k.*
2. Initialize *k* cluster centers can be done in various ways, most often done by random, cluster centers are given initial values with random numbers.

3. Place each data / object to the cluster closest proximity of the two objects is determined based on the distance between the two objects. Similarly, the proximity of the data to a particular cluster is determined by the distance between the data center cluster. In this stage, the data needs to be calculated by the distance to each cluster center. The closest distance between the data with a particular cluster will determine the data in the where cluster's.
4. Recalculate cluster center membership now with Cluster Center is an average of all the data / object in a particular cluster. If desired can also use the median of the cluster. So the average (mean) is not the only measure that can be used.
5. Assign each object again using the new cluster center. If the cluster center has not changed yet, the clustering process is complete. Or back again to the number three to cluster centers in the data could not change anymore.

As it has been said that the mean as the center of clusters can be replaced with other measures such as the median concentration. For certain cases as an alternative to the use of the median of the mean gives better results. With other words, the median is not sensitive to data outliers. Results with clusters using *k*-means clustering method depends on the value of the initial cluster centers are given. Giving different initial values will yield different clusters. There are several ways to give initial one is to take the initial sample of data and then look for its center; members with random initial value and then specify the initial value or use the results of the cluster hierarchy with an appropriate number of clusters

III. APPLICATION

A. Application data with K-means Cluster.

The purpose of this research is how to determine the three cluster for predictin of the industrial area that will be built in the Industrial Estate. It should be bult the industry area for medium enterprises which has good facilities, but to allocate industries to be in all the three types of industry area, namely: the seafood processing industry, wood processing industry and basic chemical industry, plus food. The variable that will be taken as the basis for determining the acreage is each business unit and the amount of investment that will be done; there are hundreds more units that will be made cluster prediction as follows:

Table 1. Data area and investment of various industries.

No	Area (m2) x100	No	Area (m2) x100	No	Invest (USD) x1000	No	Invest (USD) x1000
	1		2		3		4
1	2,5377	54	0.9311	1	2.2696	54	200.5784
2	3,8339	55	1.1905	2	2.4943	55	207.5784
3	0,2588	56	3.4897	3	2.0226	56	207.5784
4	2,8622	57	3.409	4	2.8622	57	207.5784
5	2,3188	58	3.4172	5	4.7694	58	203.5784
6	0,6923	59	2.6715	6	5.5784	59	250.5784
7	1.5664	60	0.7925	7	6.5784	60	255.5784
8	2.3426	61	3.4897	8	15.5684	61	260.5784
9	5.5784	62	3.409	9	16.5786	62	275.5784
10	4.7694	63	3.4172	10	17.5787	63	270.5784
11	0.6501	64	2.6715	11	15.5783	64	275.5784
12	5.0349	65	0.7925	12	16.5784	65	276.5784
13	2.7254	66	2.7172	13	23.5784	66	274.5784
14	1.9369	67	3.6302	14	14.5784	67	272.5784
15	2.7147	68	2.4889	15	18.5784	68	207.5784
16	1.795	69	3.0347	16	30.5784	69	307.5784
17	1.8759	70	2.7269	17	35.5784	70	307.5784
18	3.4897	71	1.6966	18	35.5784	71	309.5784
19	3.409	72	2.4889	19	32.5784	72	305.5784
20	3.4172	73	3.0347	20	6.5784	73	307.5784
21	2.6715	74	2.7269	21	7.5784	74	310.5784
22	0.7925	75	1.6966	22	8.5784	75	317.5784
23	2.7172	76	2.2939	23	9.5784	76	315.5784
24	3.6302	77	1.2127	24	15.5784	77	312.5784
25	2.4889	78	2.7269	25	10.5784	78	320.5784
26	3.0347	79	1.6966	26	7.5784	79	325.5784
27	2.7269	80	2.2939	27	76.5784	80	330.5784
28	1.6966	81	1.2127	28	75.5784	81	328.5784
29	2.2939	82	2.8884	29	85.5784	82	337.5784
30	1.2127	83	0.8529	30	95.5784	83	330.5784
31	2.8884	84	0.9311	31	45.5784	84	335.5784
32	0.8529	85	1.1905	32	75.5784	85	307.5784
33	0.9311	86	3.4897	33	95.5784	86	307.5784
34	1.1905	87	3.409	34	85.5784	87	407.5784
35	3.4897	88	3.4172	35	95.5784	88	407.5784
36	3.409	89	2.6715	36	96.5784	89	410.5784
37	3.4172	90	0.7925	37	97.5784	90	412.5784
38	2.6715	91	3.4897	38	98.5784	91	411.5784
39	0.7925	92	3.409	39	95.5784	92	415.5784
40	2.7172	93	3.4172	40	105.5784	93	417.5784
41	3.6302	94	2.6715	41	102.5784	94	420.5784
42	2.4889	95	0.7925	42	108.5784	95	425.5784
43	3.0347	96	2.7172	43	109.5784	96	430.5784
44	2.7269	97	3.6302	44	110.5784	97	411.5784
45	1.6966	98	2.4889	45	111.5784	98	415.5784
46	2.4889	99	2.7269	46	112.5784	99	417.5784
47	3.0347	100	1.6966	47	120.5784	100	420.5784
48	2.7269	101	2.2939	48	115.5784	101	425.5784
49	1.6966	102	1.2127	49	112.5784	102	430.5784
50	2.2939	103	2.8884	50	114.5784	103	411.5784
51	1.2127	104	0.8529	51	113.5784	104	415.5784
52	2.8884	105	0.9311	52	115.5784	105	407.5784
53	0.8529	106	1.1905	53	205.5784	106	207.5784

Source: data planning investment province

From the above data would make into three clusters, then plotted by using the high-level programming language and produced the following results:

.

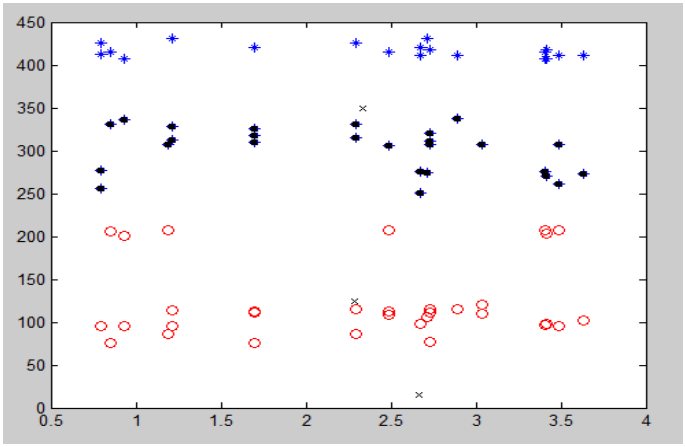

Fig.2. the three clusters for the area based on investment.

From the Figure that shown there are three clusters has divided from 106 data of medium enterprises industries. There are around 26 medium enterprises in cluster three, 49 stand in to the second cluster and 31 in the first cluster. The *silhouetted* value of that data is:

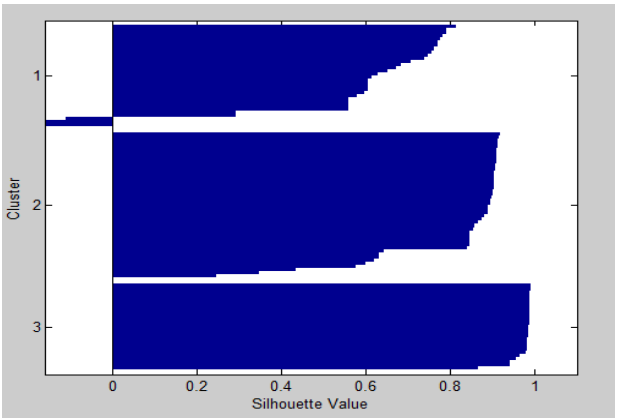

Fig.4. the three clusters for the area based on investment.

The performance of the data can show that there is the longer data belong to the three cluster with h value is 1.

B. Application data with Hierarchical Cluster

In the hierarchical clustering the distances calculated the data with each other and the data will be obtained vector distance from each object to another object, a tree can be determined by finding the pair nearest cluster of each object or group of objects. Figure cluster hierarchy can be seen as follows:

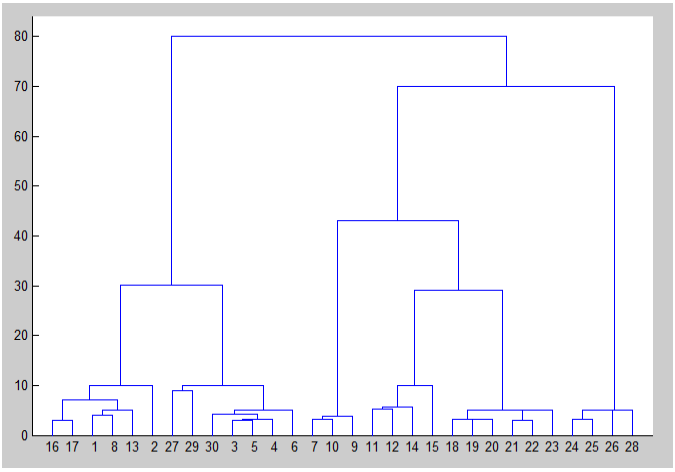

Fig.3. the result of data in hierarchy clusters.

The command will be used to generate linkage grouping, then using the command from the command linkage dendrogram result will generate a cluster hierarchy representation. In this image will be described a linear relationship between the correlation coefficient obtained *cophenetic* distances from the cluster tree computes the mean coefficient how good dissimilarity between the data where the distance between two data *cophenitic* represented by a dendrogram by the height where the two lines connecting the data put together the first time. The height is the distance between the two sub-clusters united by a hyphen

IV. EXPERIMENTAL RESULT

From the process, produce result different cutting cluster points, for K-means show that the cutting point in first cluster is in 31 while for the hierarchy is 52, and for the second cluster one method is 80 and the hierarchy is 86, it shown that there is a little bit different on the cutting point, but generally almost the same. This is summary as follow:

Tabel.2. Result of the two methods

Cluster	K-means		Hierarchy	
I	31	1-31	52	1-52
II	49	32-80	34	53-86
III	26	81-106	20	87-106

In planning cluster first should be for the seafood processing industry the second cluster for wood processing industry and the third for basic chemical industry plus food. For a planner must be manage the condition where the mainly kind of sea food processing, wood processing industry and basic chemical industry plus food where kind of unit able to stand up alone and where is can be nearer with other industry cluster belong them.

V. CONCLUSION

In summary, this paper introduces two methods-clustering algorithms to solve clustering problems. First determine the number of clusters. Then with the software process to get the results of the optimization algorithm is applied to find a good final solution. In principle, the K-means more showed a little bit better results, but the method of approaching the desire clustering offer decision area data that some are in a transition area when compared to the K-mean, it was an interesting thing because there are real-industrial could adjacent, though not of a type of industry, but not interfere with each other and affect or it can be interconnected in terms of the product chain. Such as K-means clustering algorithm which has the simplicity and high speed in clustering large datasets? Based on the results of K-means algorithm and the hierarchy in the application to find the final solution of the clustering is to have the good of each, but they can be equated, so that the second method is a robust clustering method. It can be applied as a new clustering method for grouping other problems.

REFERENCES

[1] A.Kocsor, D., Laszlo Toth and Denes Paczolay," A nonlinearized discriminant analysis and its application to speech impediment therapy", Proceedings of Text, Speech and Dialogue: 4th International Conference, TSD 2001, LNAI 2166,(ed) V.Matousek, R. M. K. T., P. Mautnoceer, Spring Verlag, pp.249-257, 2001.

[2] Bhattacharyya, C., Grate, L.R., Rizki, A., Radisky, D.C., Molina, F.J, Jordan, M.I.,Bissell, M,J. and Mian, I.S."Simultanieous relevant feature identification and classification in high dimention spaces: application to molecular profiling data." Signal Processing (2002)

[3] C. C Chang and C.J Lin, LIBSVM: A Library for Support Vector Machines, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] C.W. Hsu and C.-J.Lin A Simple Decomposition Method for Support Vector Machines, machine Learniong 46(2002), 291-314.

[5] C.W.Hsu and C.-J. Lin, A Comparison of Method for Multi-class Support Vector Machines, IEEE Transaction on Neural Networks, 13(2002), 415-425.

[6] Dolores M. Etter and David C.Kuncicky, Introduction to Mathlab 7, with Holly Moore. Up-per Saddle River, NJ: Pearson/Prentice Hall, 2005.

[7] Girolami, M., 2002,"Mercer kernel-based clustering in feature space", IEEE Trans Neural Networks, 13, no.3, 780-784.

[8] G.L Press, 2002.

[9] Hastie, T., Tibshirani, R., and Friedman,J., 2001, The Elements of Statistical Learning: data mining, inference, and prediction, Springer-Verlag, New york.

[10] Heijden, F., Duin, R.P.W., Ridder, D, Tsx, D.M.J., Classification, Parameter Estimation and State Estimation: An Engineering Approch Using MATHLAB, john Wieley and Sons Ltd., West Sussex, England, 2004.

[11] Junshui Ma, Yi Zhao, and Stanly Ahalt,"OSU SVM Classifier Mathlab Toolbox", available at http://www.kernel-machines.org/

[12] Richard O.Duda, Peter E. hart, David G.Stork, Pattern classification, john wiley and Sons, New york 2001.

[13] Santosa, Budi, Data Mining: Teknik Pemanfaatan data Untuk Bisnis, Teori dan Aplikasi, Graha Ilmu, 2007.

[14] S.R.Gunn, Technical report, Dept.of electronics and Computer Science support Vector Machines for classification and Regression, University of soutampton, (Southamton, U.K), 1998.

[15] S.Mika, G.Ratsch, and K-R.Muler. A mathematical programming approach to the Kernel Fisher algorithm. In T.K.Lean, T.G.Dietterich, and V.Tresp, editors, Advances in Neural Information Processing Systems 13, Pages 591-597. MIT Press, 2001.